Joint EGEE/OSG VO Management at HPDC '08

GlideinWMS

The CMS pilot infrastructure

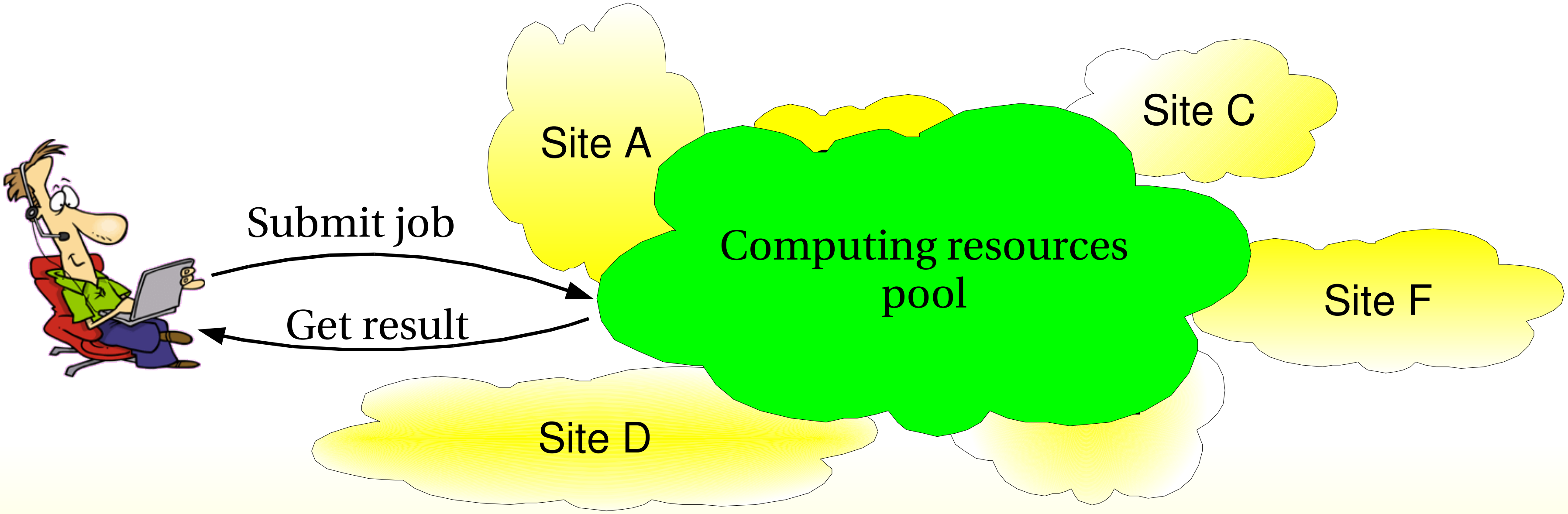
by Igor Sfiligoi (Fermilab)

Outlook

- Grid computing overview
- The pilot paradigm
- Introducing Condor glideins
- glideinWMS description
- glideinWMS in real life
- Conclusions

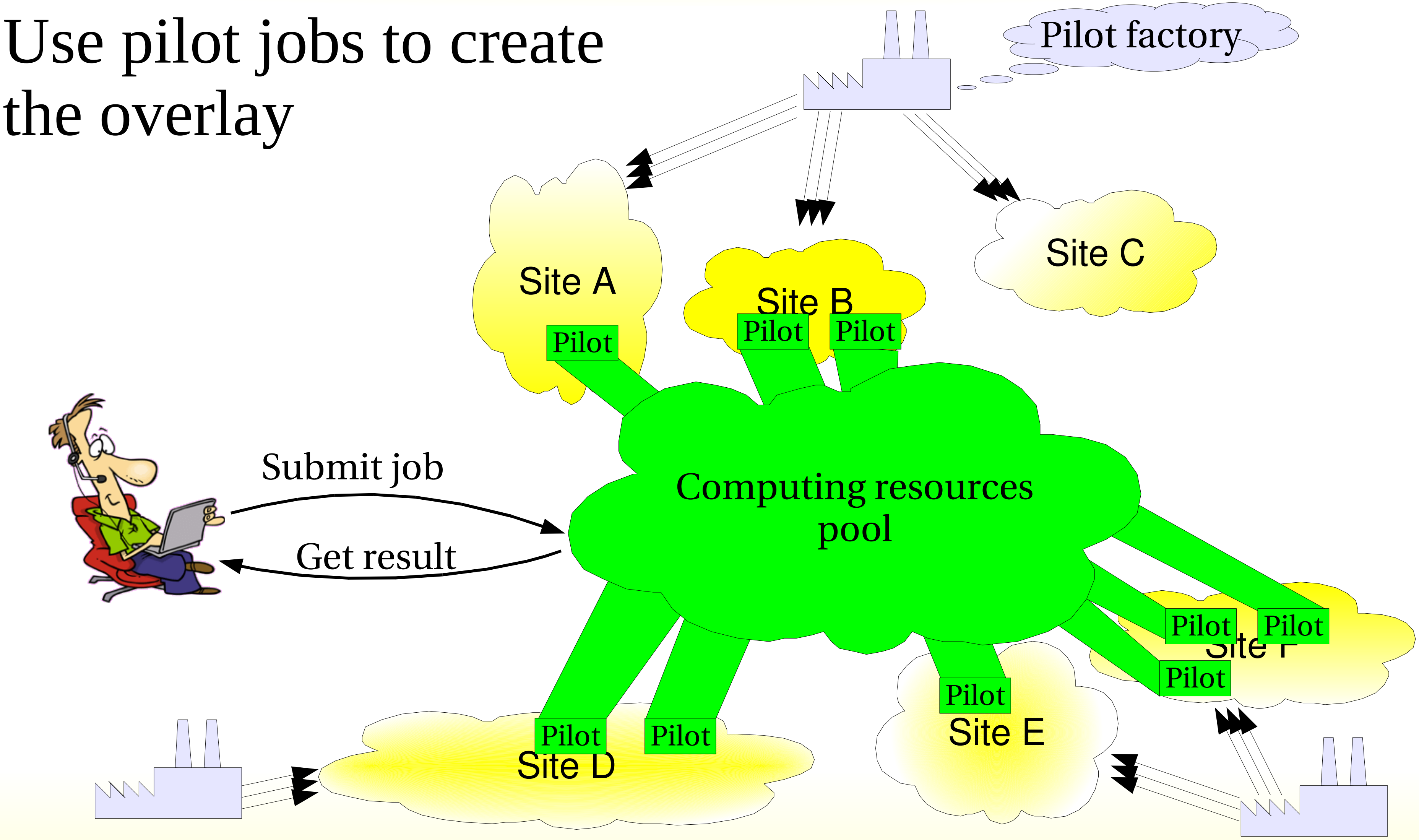
Bare-bones Grid is complex

- The problem
  - The Grid is a heterogeneous set of computing sites
    - Deciding where to run a job is far from trivial
- Possible solution
  - Make the grid uniform by creating an overly layer



The pilot paradigm

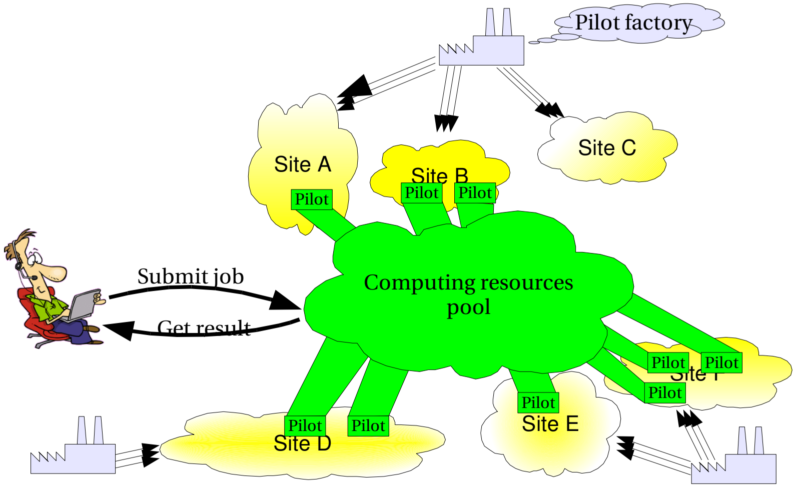
- Use pilot jobs to create the overlay



The pilot paradigm (continued)

- Use pilot jobs to create the overlay
  - Never send user jobs directly
- When a pilot lands on a Grid worker node
  - Validates Grid resource
  - Prepares the environment
  - Pulls a user job
- Hides Grid heterogeneity
  - Users see a fairly uniform computing pool

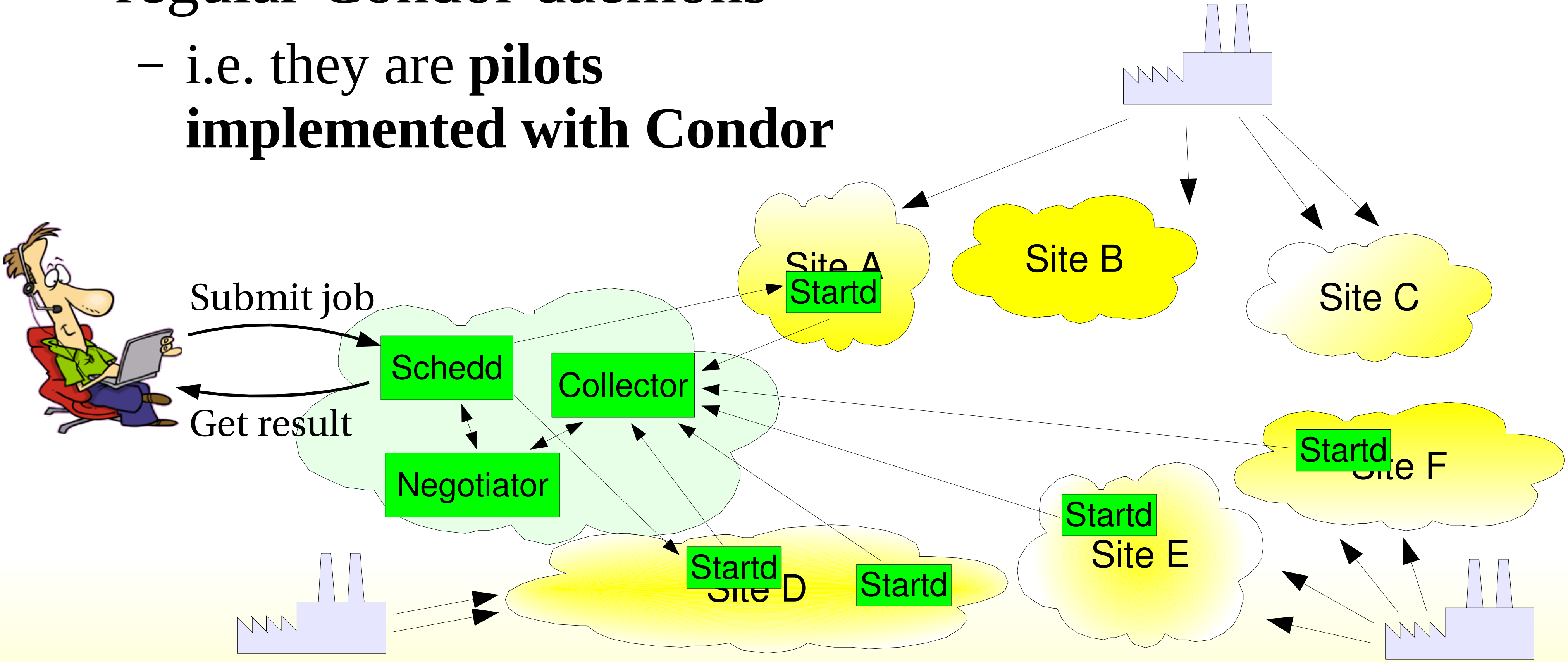
User jobs
never sent directly
to Grid sites



Condor glideins

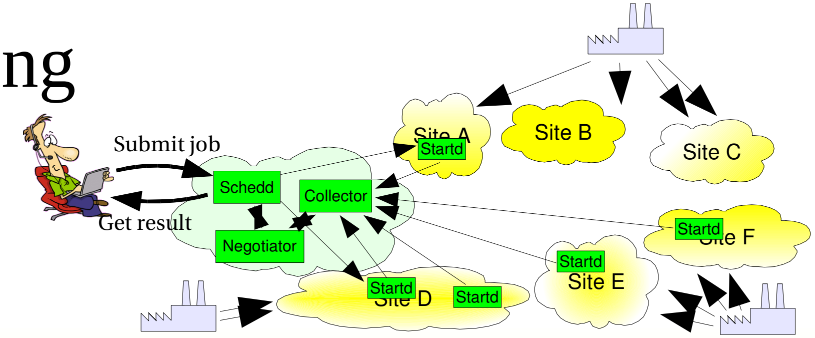
http://www.cs.wisc.edu/condor/

- Condor is based on a distributed architecture
- Condor glideins are Grid jobs that start regular Condor daemons
  - i.e. they are **pilots implemented with Condor**



Submitting glideins

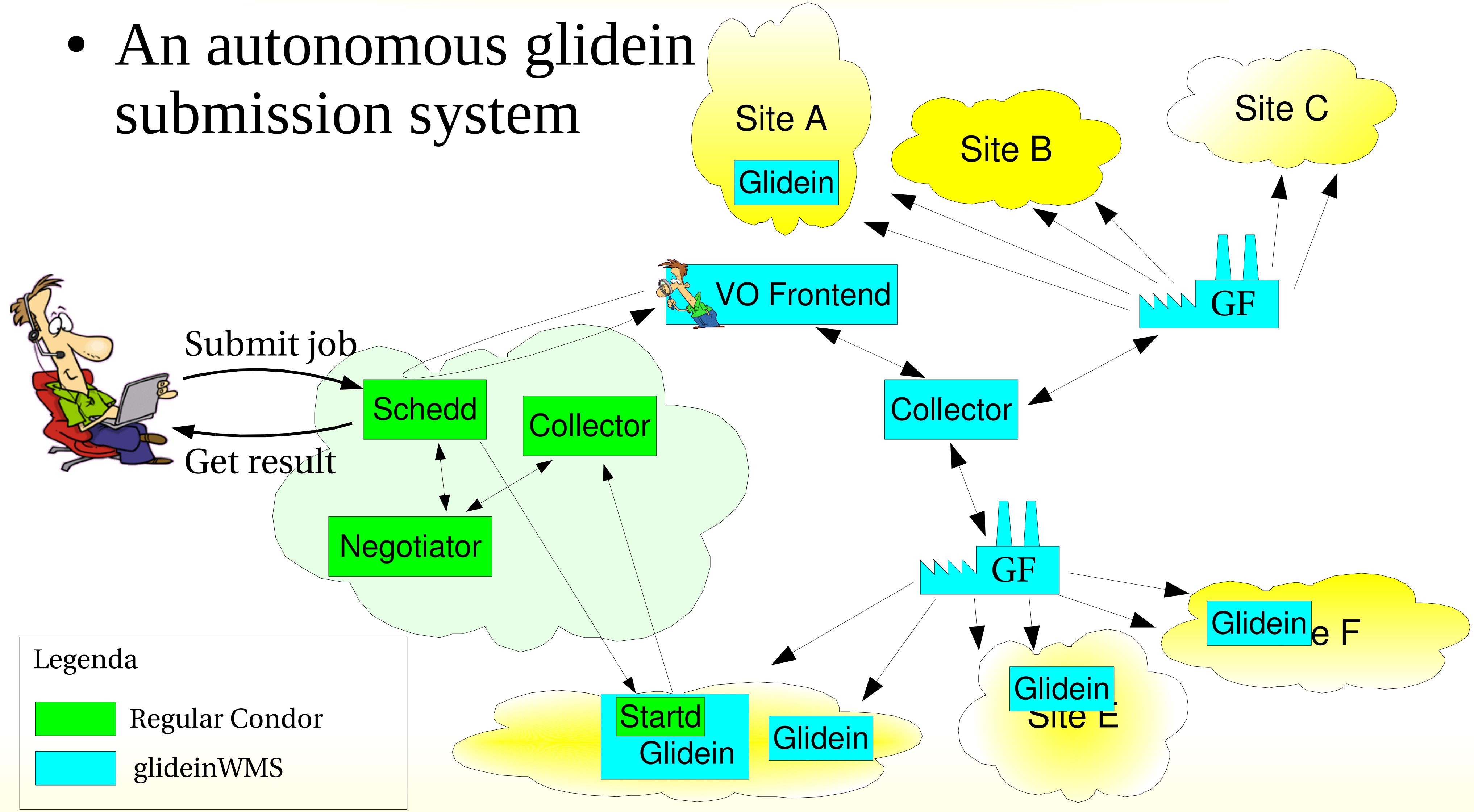
- Condor provides only a basic command line glidein submission tool
  - Good for trying out glideins
  - But not meant to be used as a glidein factory
- A few groups developed glidein factories
  - CDF has the CDF-specific GlideCAF
  - USCMS@FNAL is developing the **glideinWMS**



Introducing the glideinWMS

http://www.uscms.org/SoftwareComputing/Grid/WMS/glideinWMS/

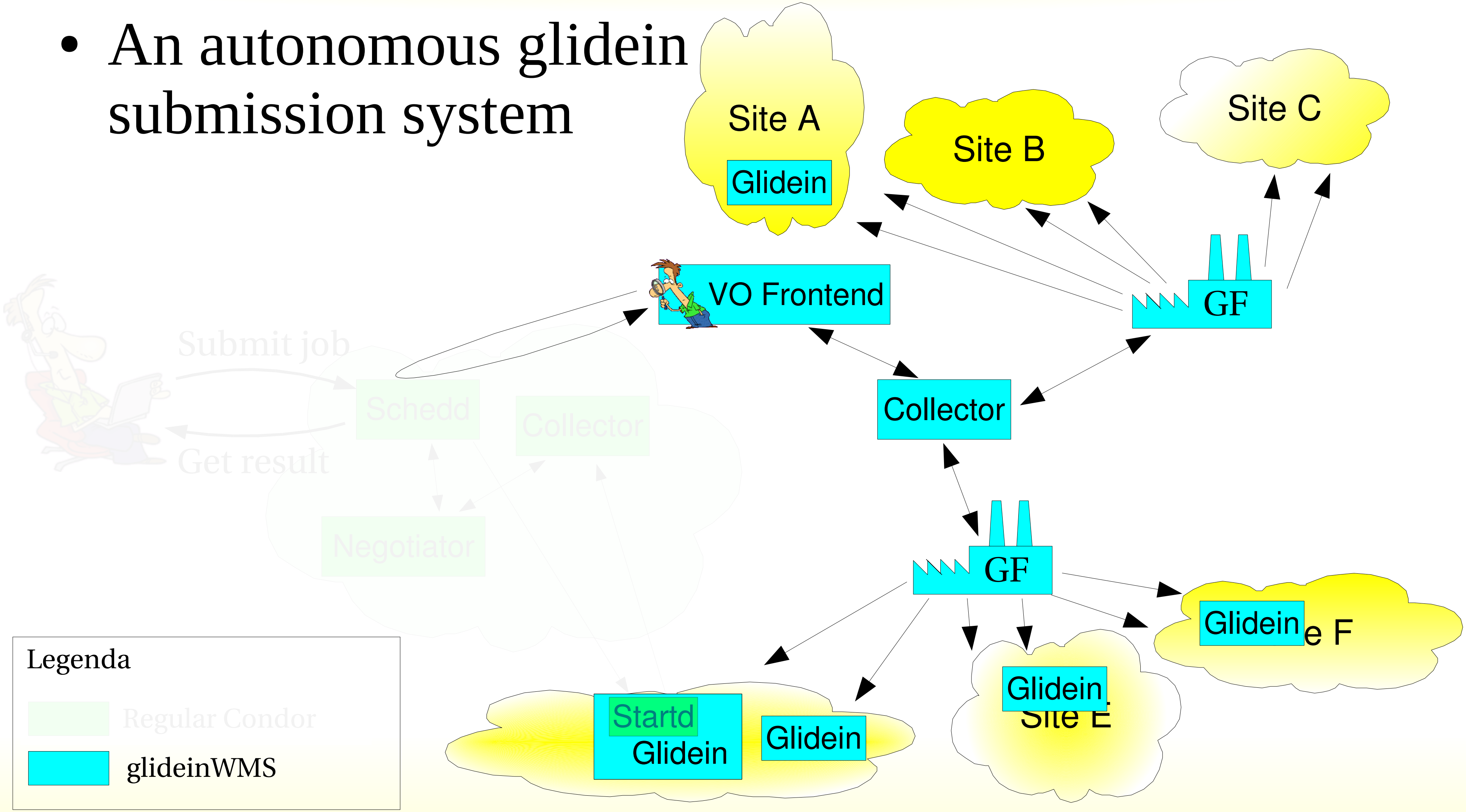
- An autonomous glidein submission system



Introducing the glideinWMS (2)

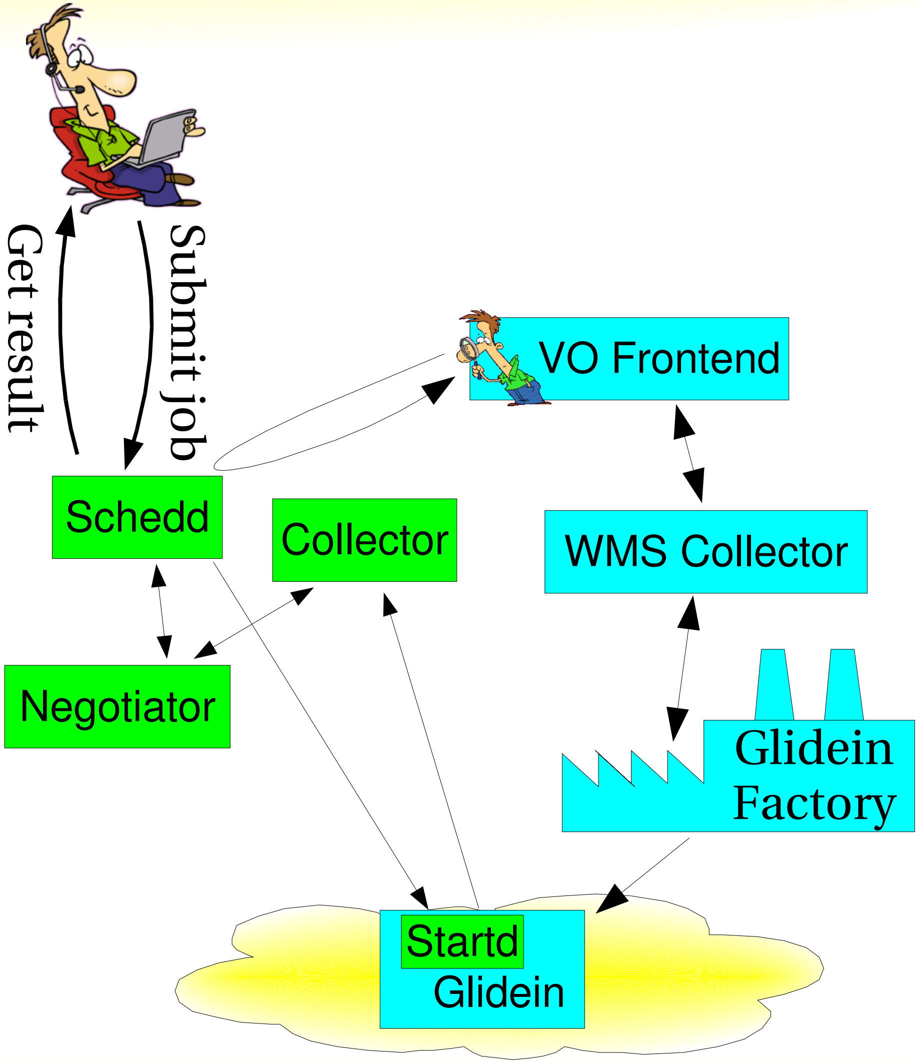
http://www.uscms.org/SoftwareComputing/Grid/WMS/glideinWMS/

- An autonomous glidein submission system



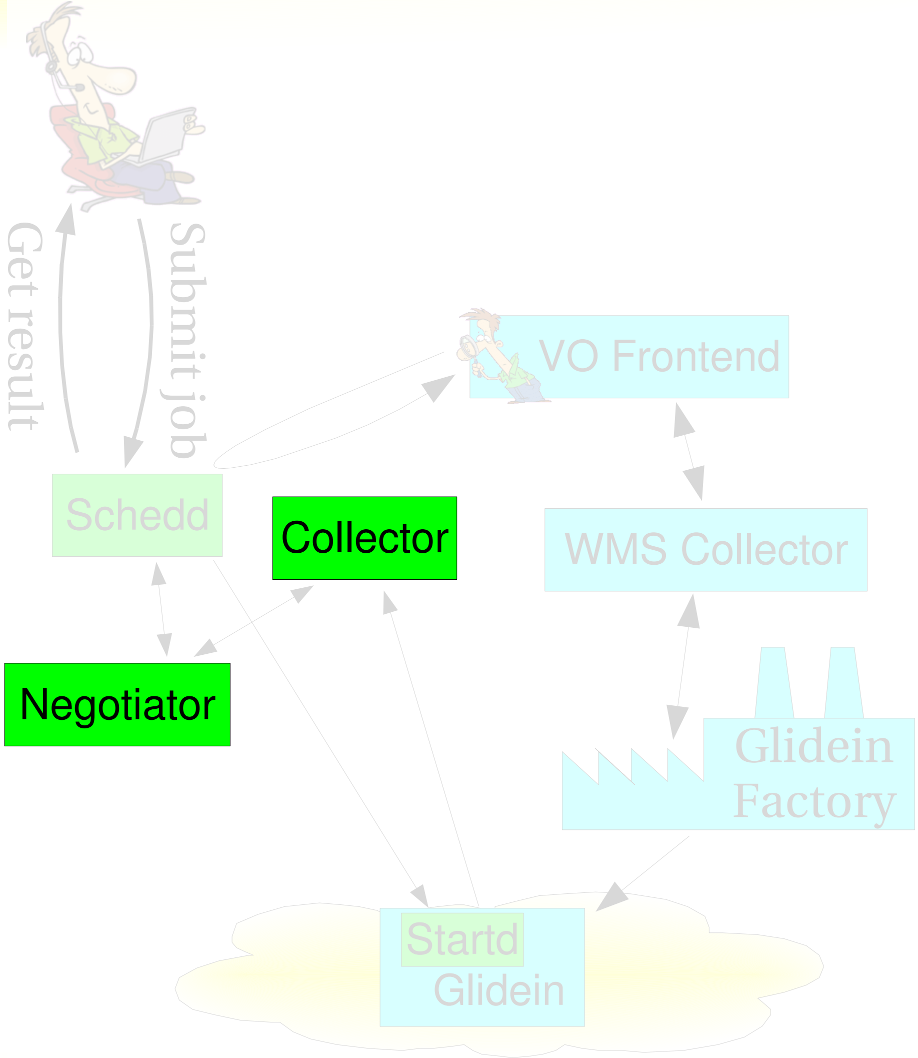
glideinWMS architecture

- glideinWMS composed of six logical pieces:
  - A Condor central manager (collector + negotiator)
  - One or more Condor submit machines
  - A glideinWMS collector
  - One or more VO frontends
  - One or more glidein factories
  - The glideins



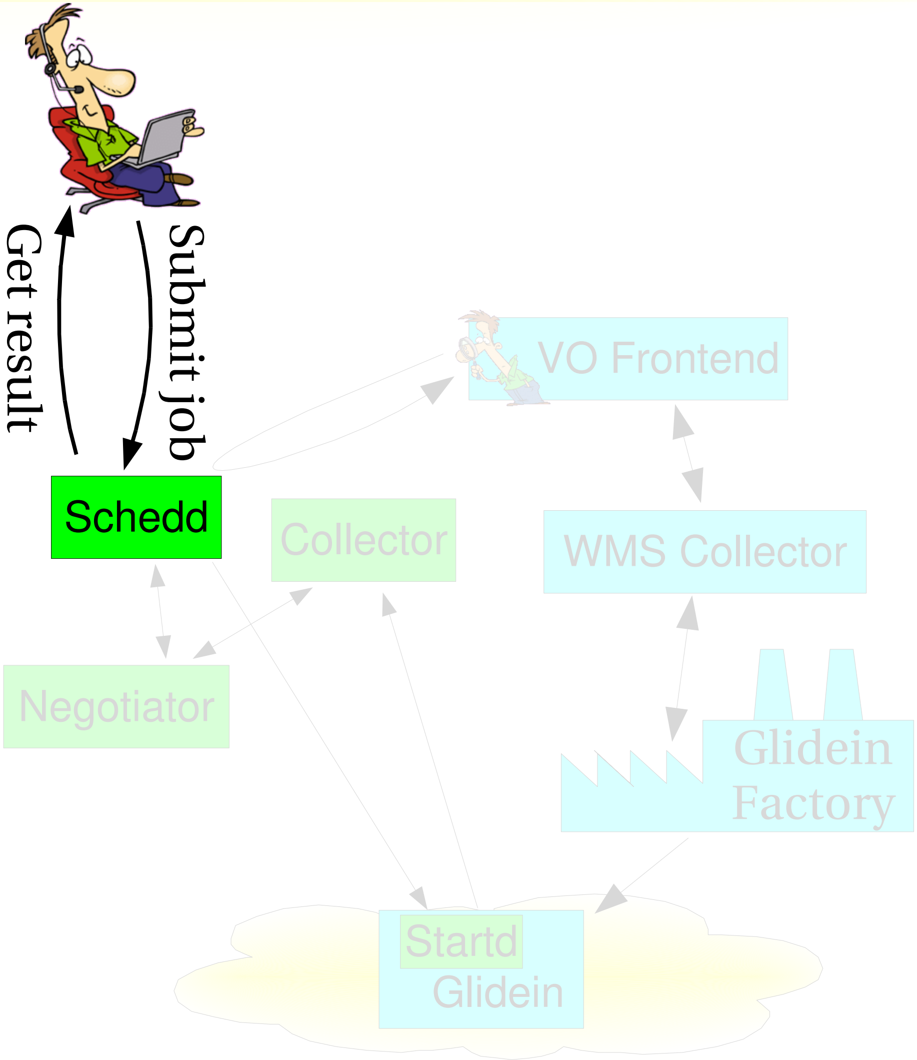
glideinWMS architecture (2)

- glideinWMS composed of six logical pieces:
  - A Condor central manager (collector + negotiator)
  - One or more Condor submit machines
  - A glideinWMS collector
  - One or more VO frontends
  - One or more glidein factories
  - The glideins



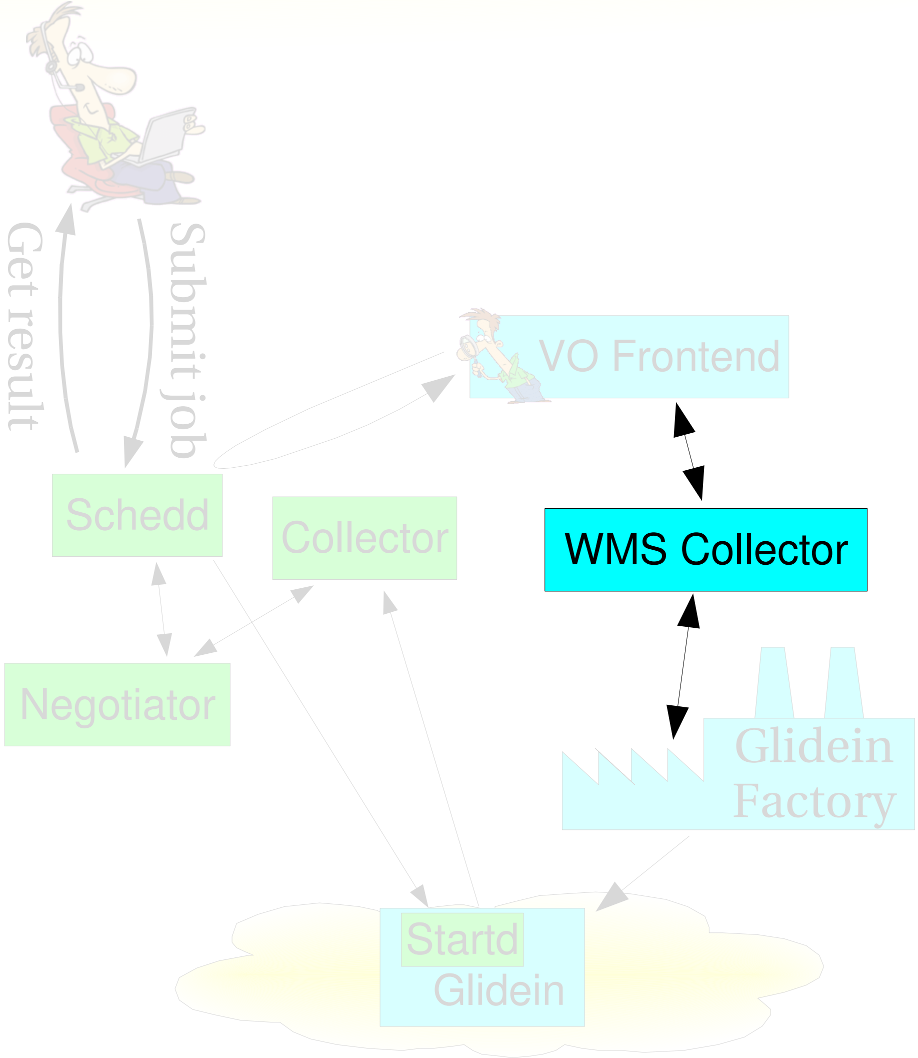
glideinWMS architecture (3)

- glideinWMS composed of six logical pieces:
  - A Condor central manager (collector + negotiator)
  - One or more Condor submit machines
  - A glideinWMS collector
  - One or more VO frontends
  - One or more glidein factories
  - The glideins



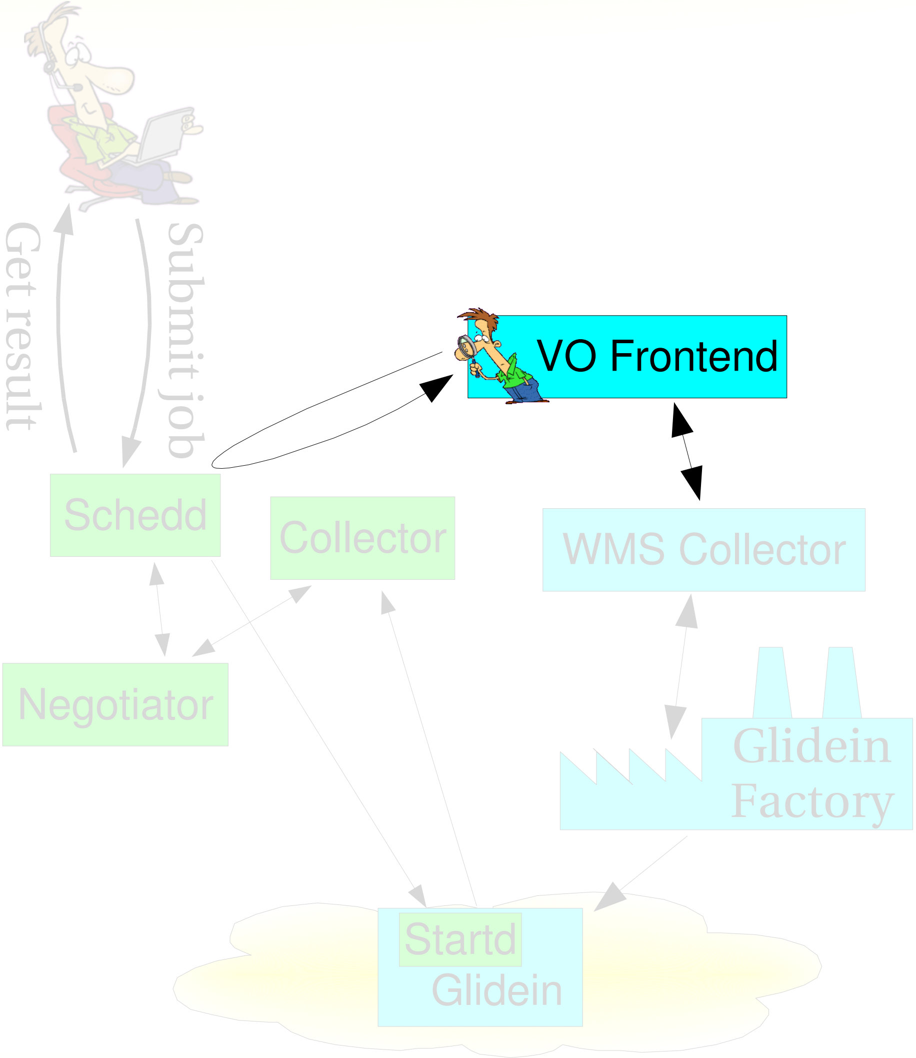
glideinWMS architecture (4)

- glideinWMS composed of six logical pieces:
  - A Condor central manager (collector + negotiator)
  - One or more Condor submit machines
  - A glideinWMS collector
  - One or more VO frontends
  - One or more glidein factories
  - The glideins



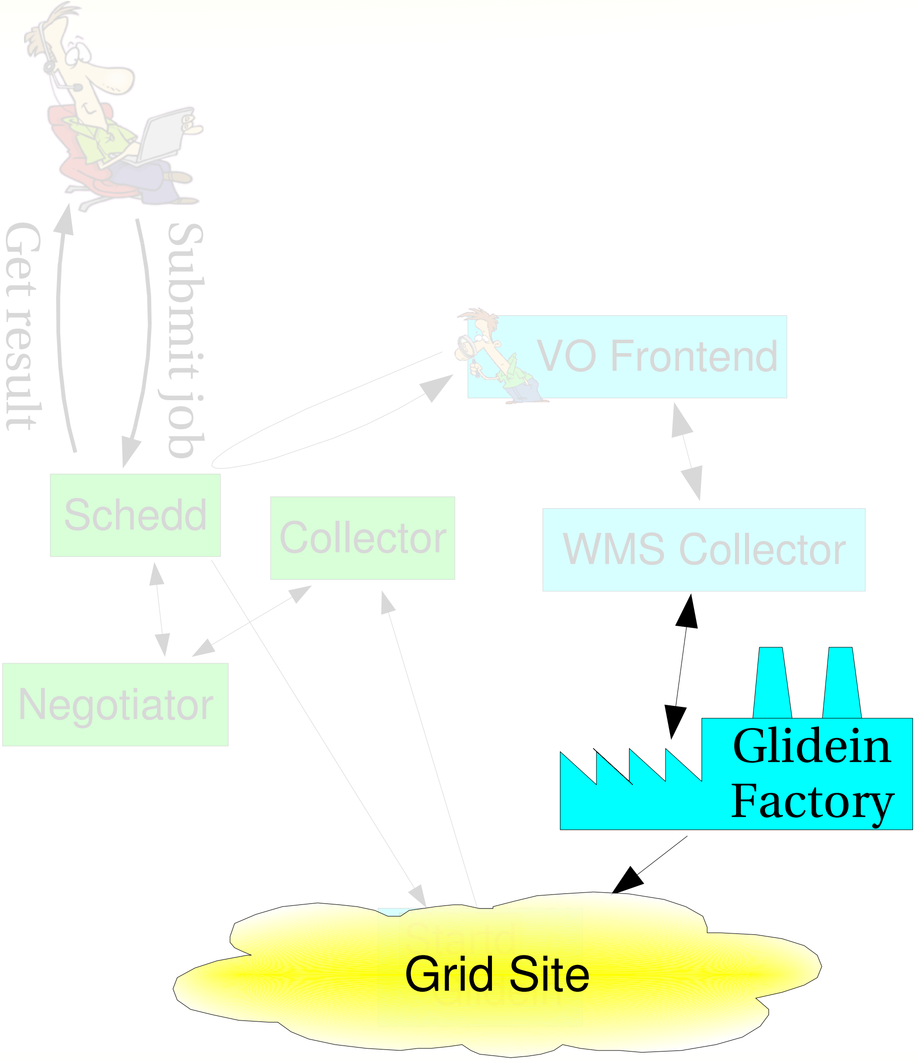
glideinWMS architecture (5)

- glideinWMS composed of six logical pieces:
  - A Condor central manager (collector + negotiator)
  - One or more Condor submit machines
  - A glideinWMS collector
  - One or more VO frontends
  - One or more glidein factories
  - The glideins



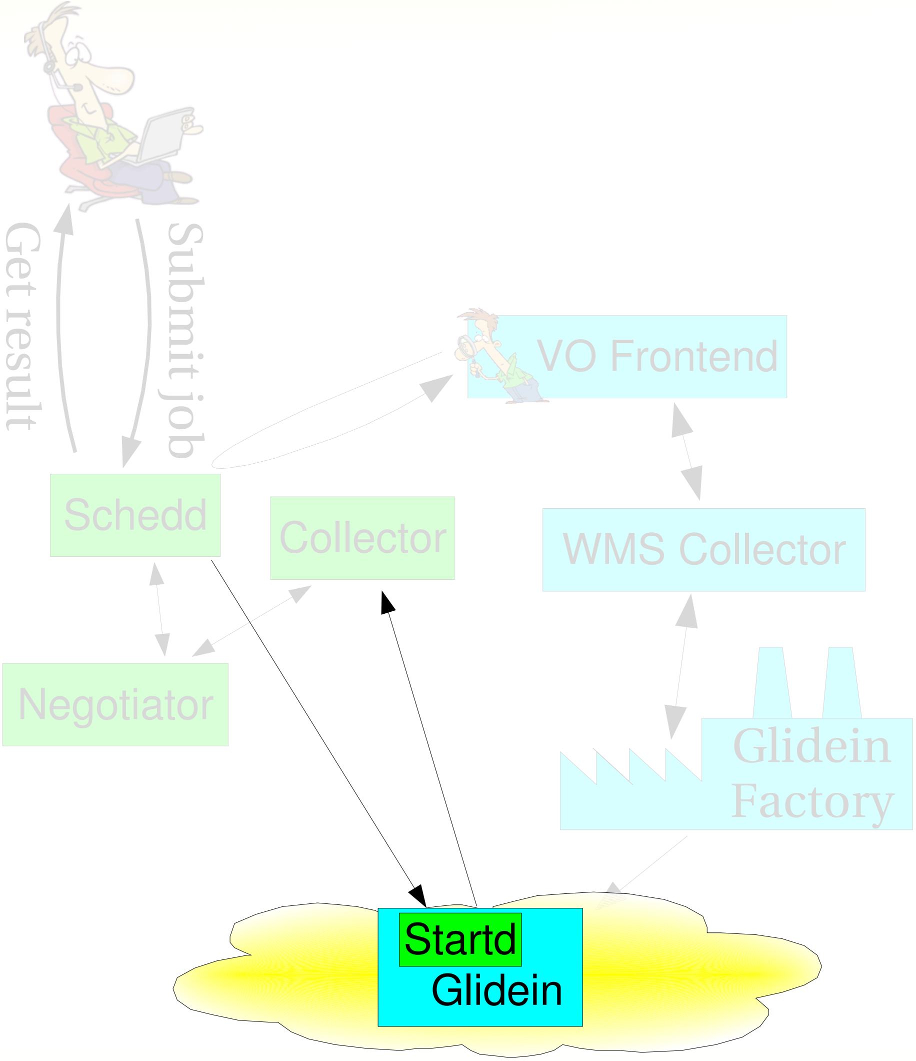
glideinWMS architecture (6)

- glideinWMS composed of six logical pieces:
  - A Condor central manager (collector + negotiator)
  - One or more Condor submit machines
  - A glideinWMS collector
  - One or more VO frontends
  - One or more glidein factories
  - The glideins



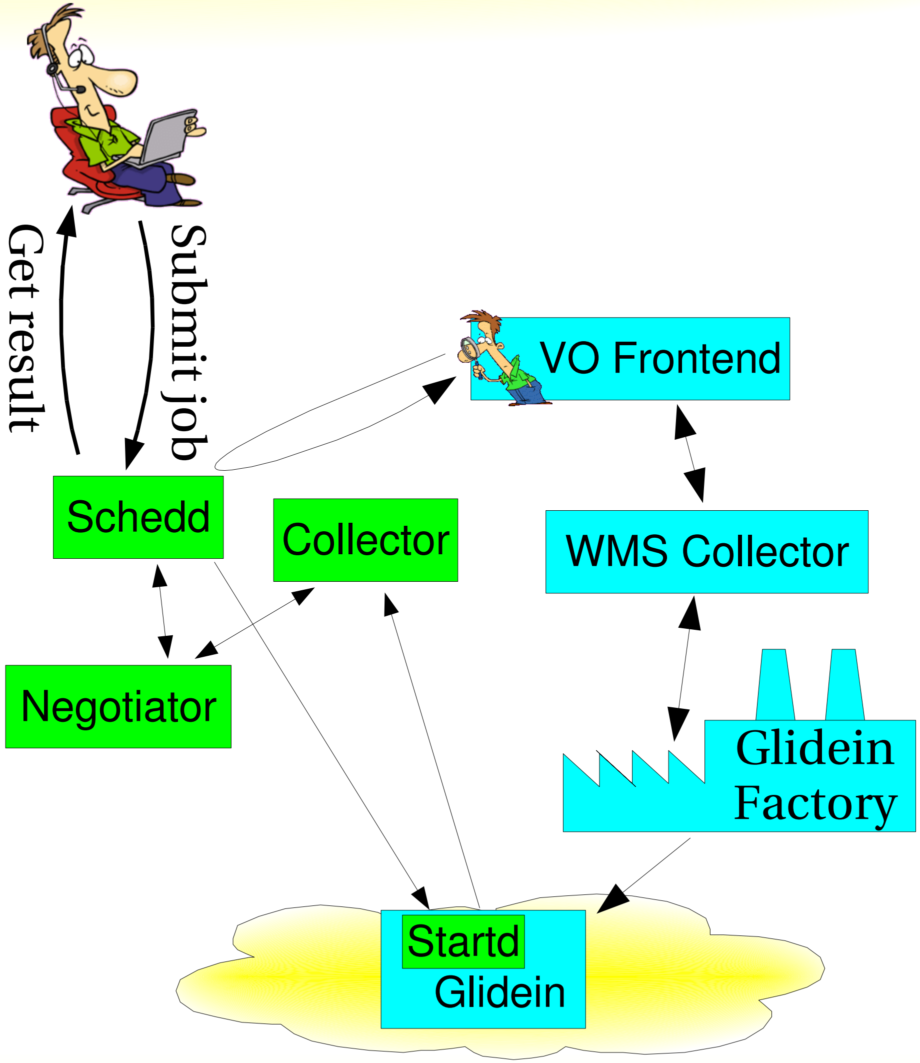
glideinWMS architecture (7)

- glideinWMS composed of six logical pieces:
  - A Condor central manager (collector + negotiator)
  - One or more Condor submit machines
  - A glideinWMS collector
  - One or more VO frontends
  - One or more glidein factories
  - The glideins



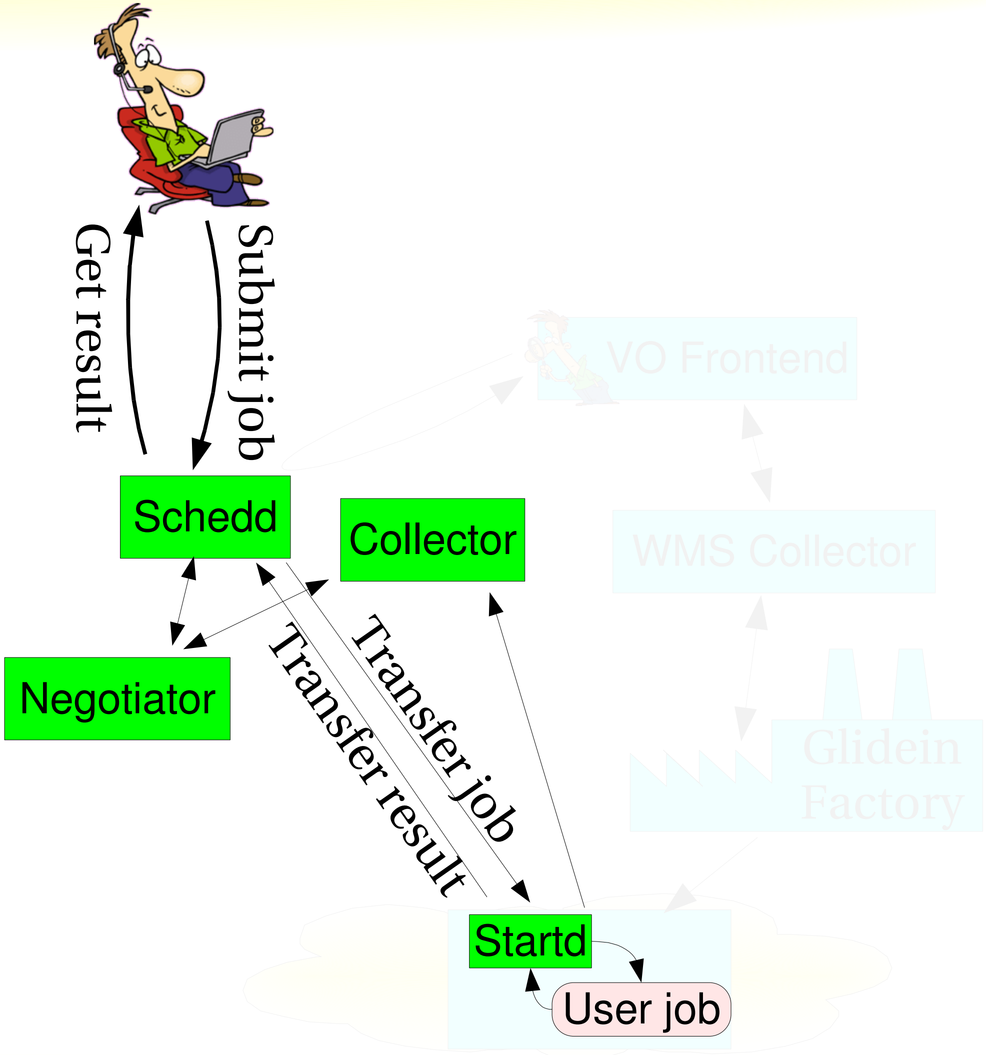
glideinWMS architecture (8)

- glideinWMS composed of six logical pieces:
  - A Condor central manager (collector + negotiator)
  - One or more Condor submit machines
  - A glideinWMS collector
  - One or more VO frontends
  - One or more glidein factories
  - The glideins



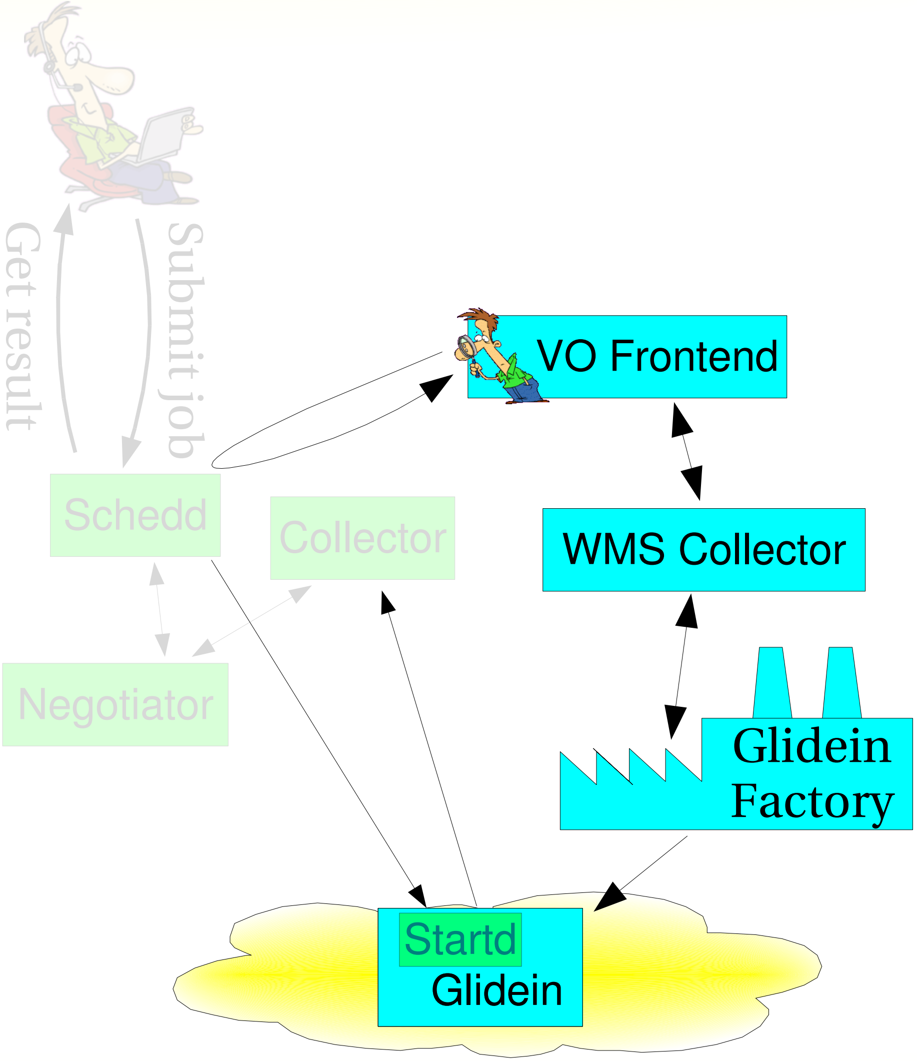
Condor handles user jobs

- A glidein Condor pool is still a Condor pool
  - Just a very dynamic one
- All Condor features available
  - ClassAds
  - Fair share
  - Group quotas
- Users really don't know about the glideinWMS



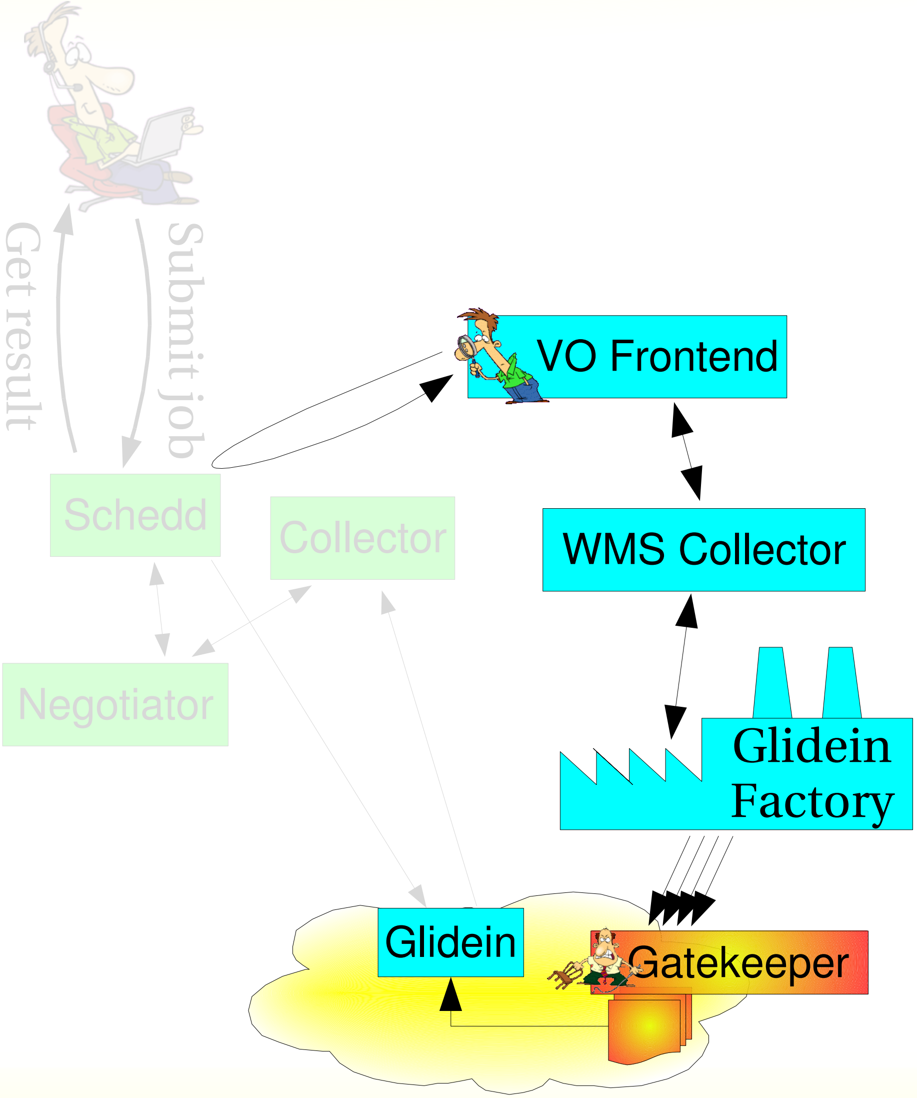
Glidein submission

- glideinWMS processes are responsible **only** for startd startup
  - A glidein just configures and starts it
  - Once started, startd has full control
- Glideins highly customizable
  - Glidein factory allows for plugins



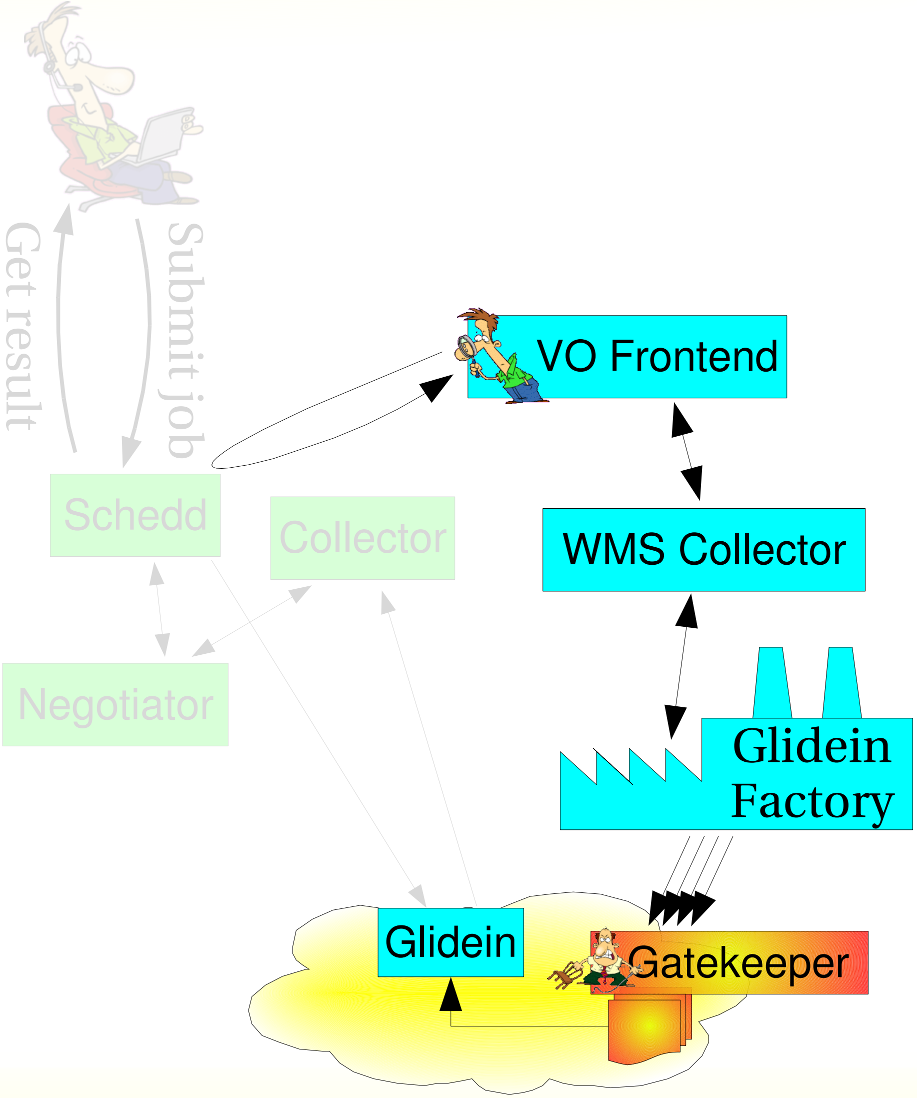
Glidein submission (2)

- Based on the principle of constant pressure
  - As long as there are enough waiting jobs in the queue, a fixed number of glideins are kept at each suitable Grid site
- Works nicely for systems with lots of waiting jobs
  - Will waste resources on seldom used systems



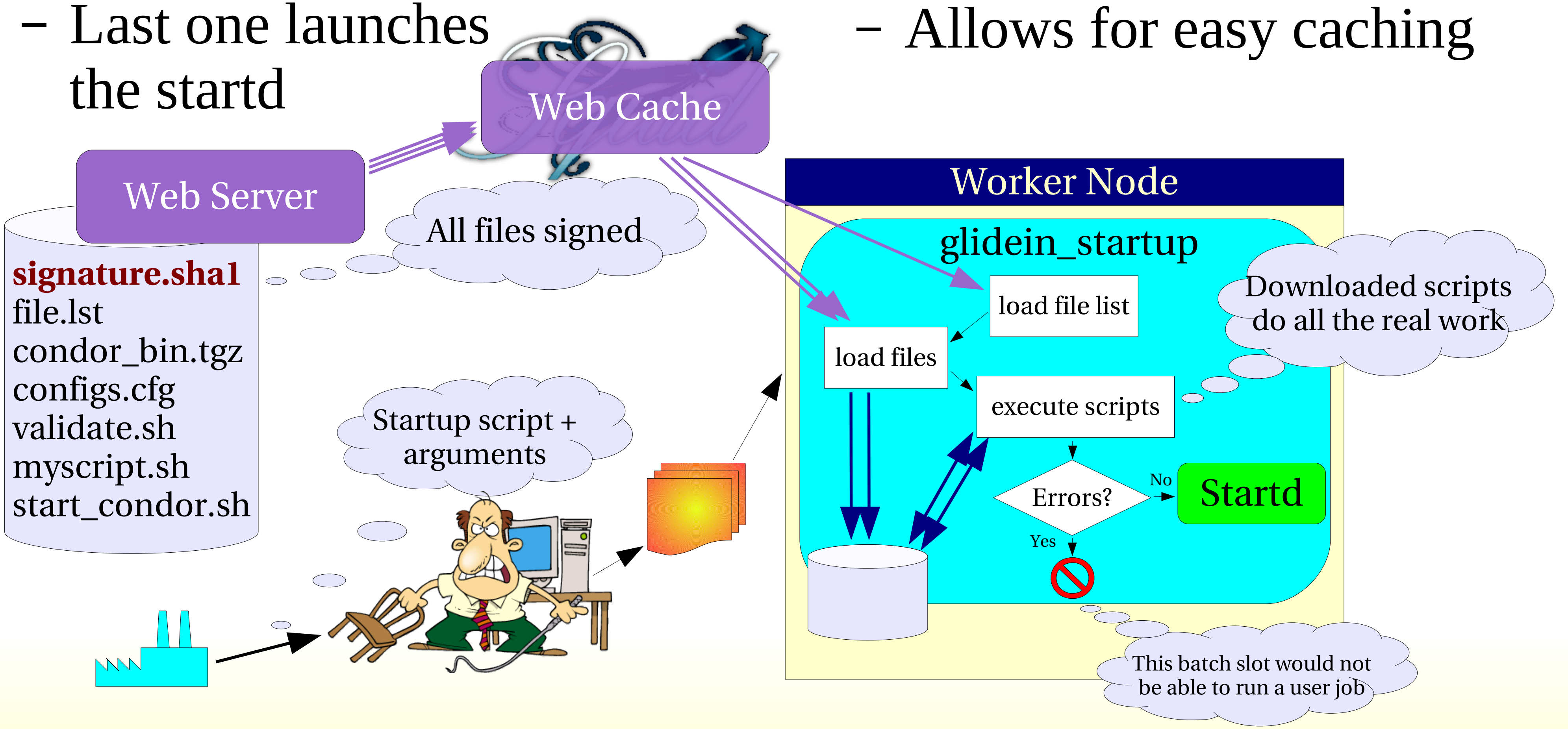
Glidein submission(3)

- Glidein submission is a collaborative work
  - VO frontend decides how many glideins to submit
  - Glidein factory actually does the submission
  - WMS collector is used for message passing
- Condor-G used for submission to Grid sites
  - Current implementation



Glidein internals

- The glidein startup script loads other plugins
  - Last one launches the startd
- HTTP used for file transfer
  - Allows for easy caching

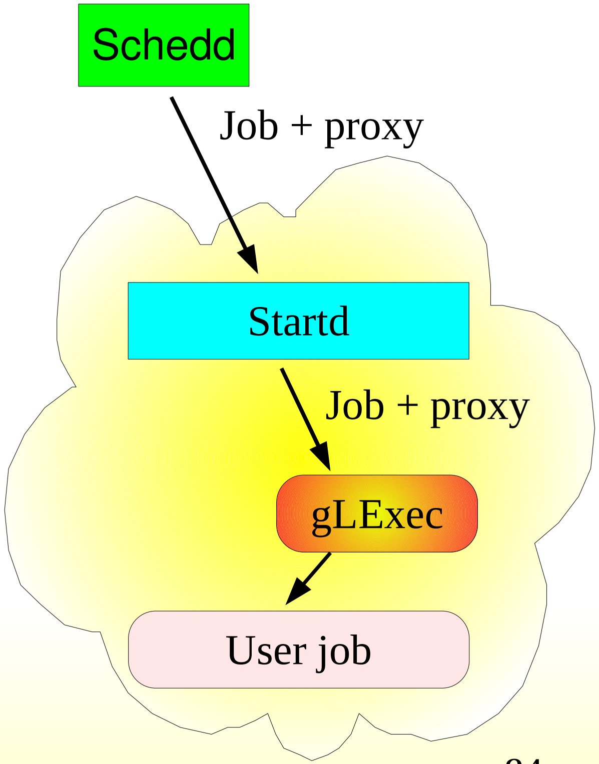


Security considerations

- GlideinWMS **requires** security over the wire
  - WAN network connections cannot be blindly trusted!
- All network traffic features integrity checks
  - Prevents man-in-the-middle attacks
- GSI authentication (X509 certificates/proxies) needed for all interactions with Condor daemons
  - Only trusted VO frontends can give orders to the glidein factories
  - Only trusted glideins can join the pool and fetch user jobs

Security considerations (2)

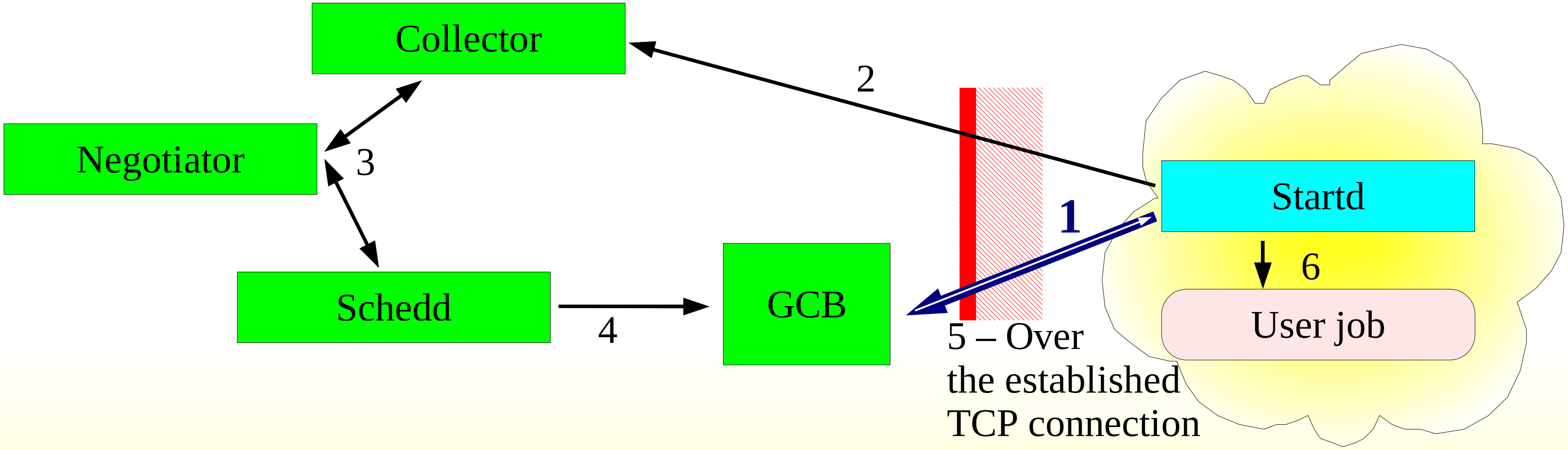
- Startd not running as a privileged user
  - Cannot change UID by itself when starting user job
  - Malicious user job could hijack the startd if running under the same UID
- Condor interfaced to gLExec
  - gLExec allows to change UID given user proxy
  - Startd protected from the user job
- gLExec part of OSG distribution
  - Deployed at several sites
  - Expected in EGEE soon



Working over the firewalls

http://www.cs.wisc.edu/condor/gcb/

- Condor uses two-way communication
  - But incoming connection often blocked by Grid sites
- Can use Condor GCB (Generic Connection Broker) to make all communications one-way
  - By opening a long lived TCP connection
  - Outgoing connectivity always needed



User job monitoring

- Good monitoring a must for most users
- Condor provides a plethora of monitoring tools
  - Most useful are condor_q and condor_status
  - Third parties provide additional Condor monitoring tools
- glideinWMS provides tools for pseudo-interactive monitoring
  - ls, cat, top on the worker nodes
- The glidein factory also maintains a basic Web based graphical view
  - plus machine readable XML and rrd data

glideinWMS monitoring

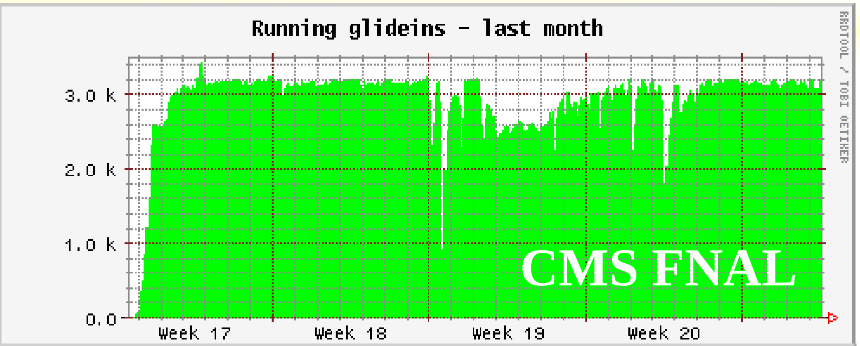
- Good monitoring a must for most administrators, too
- Condor-G provides some tools
  - Mostly condor_q
- The glidein factory maintains a rich Web based graphical view
  - plus machine readable XML and rrd data
- Glideins return comprehensive logs
  - Useful for low level debugging
  - But require some expertise to browse though

Status of glideinWMS

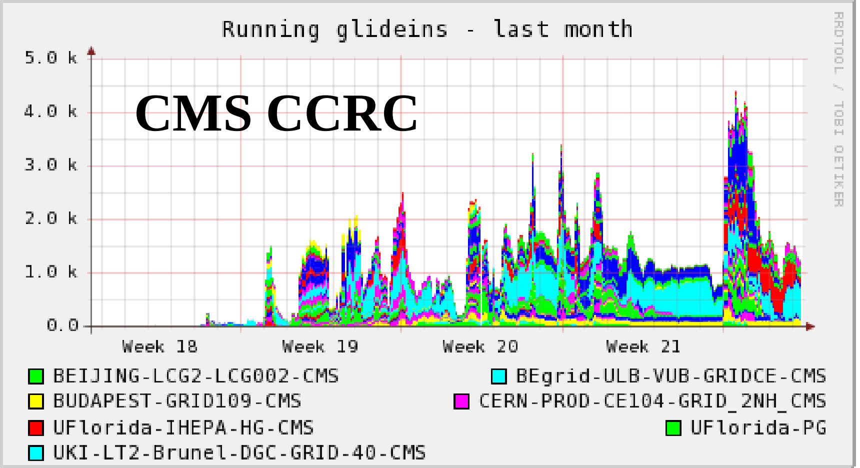
- Version 1.2.1 released May 30th
- Should be usable out of the box for most users
  - CMS is using it since v1.1
- Still in active development phase
  - More monitoring
  - More automated error checking
  - More automated error recovery
  - Better integration with other systems
- Condor also an evolving product

Glidein deployments in HEP

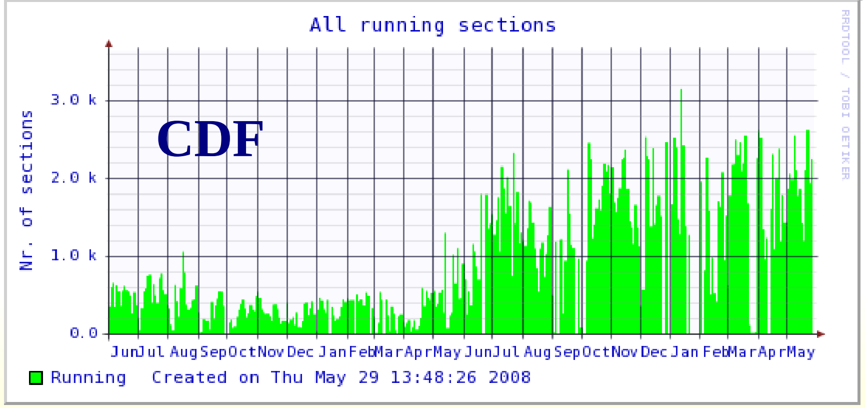
- CMS using glideins for production jobs at FNAL
  - Recently across all seven T1s
- CMS used them for analysis jobs in CCRC08
  - Across 40 T2s
- CDF and MINOS using them for user analysis



CMS FNAL



CMS CCRC



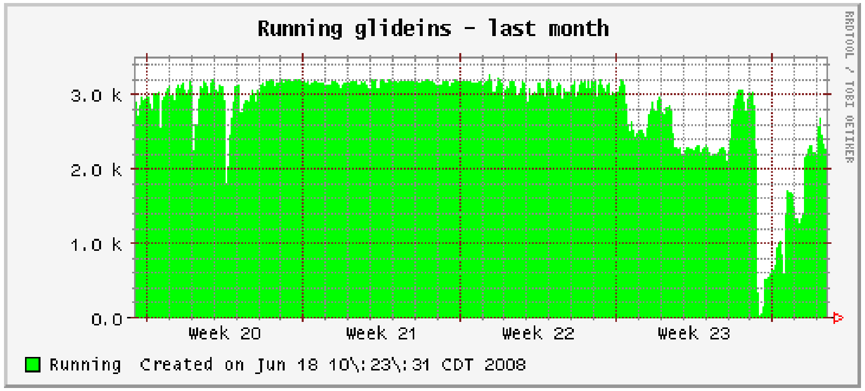
CDF

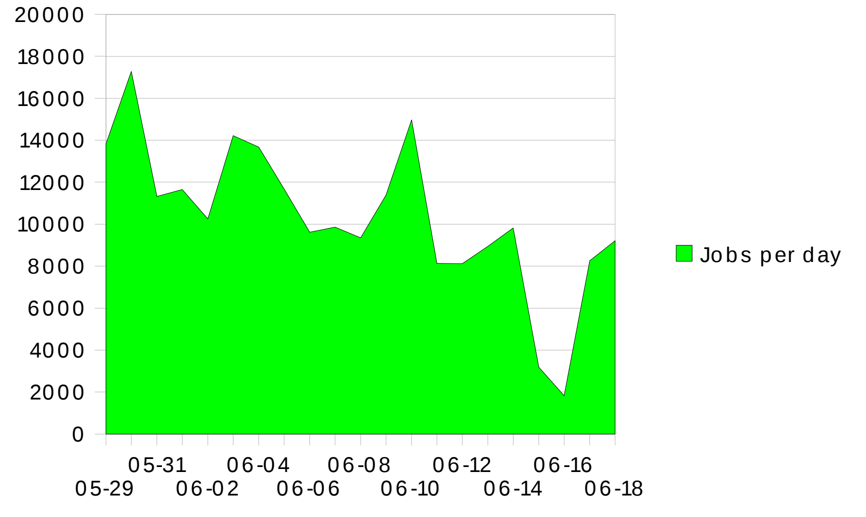
glideinWMS in numbers

- Deployed systems
  - CMS@FNAL stable 3k glideins for the past 6 months
  - CMS@CCRC up to 4k glideins over 40 sites globally
  - CDF average 2k glideins with 100s of users for past 2 years (by using the GlideCAF)
- glideinWMS Tested on a dedicated test pool, scaled without major problems to
  - 10k glideins at any time
  - 100k user jobs queued

CMS @ FNAL experience

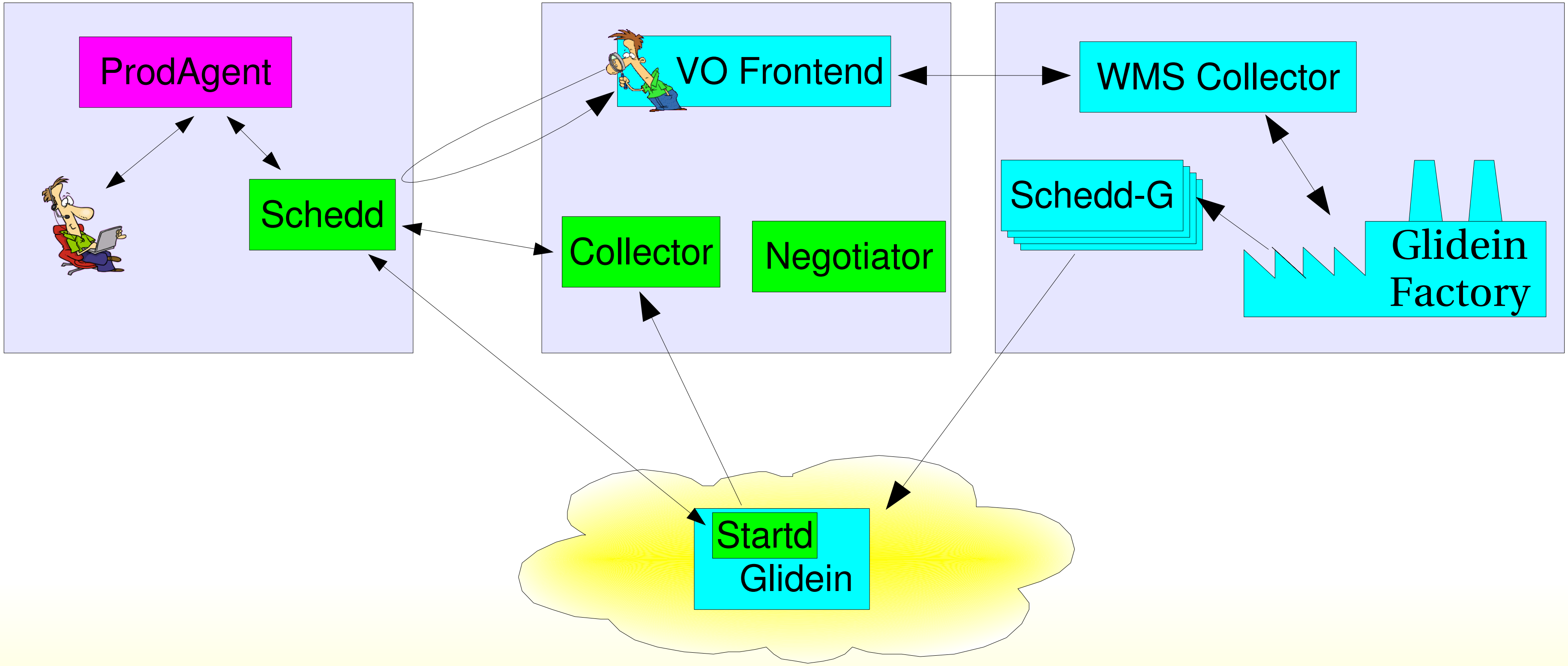
- Using ProdAgent to submit jobs to local schedd
- Gliding into a single site
  - over LAN
  - Using 3 CEs
- Saturating the FNAL T1
  - ~3200 slots
- Quick job turnaround
  - >10k jobs per day on average
  - >150k jpd during CSA07
- Few failures
  - Mostly storage related





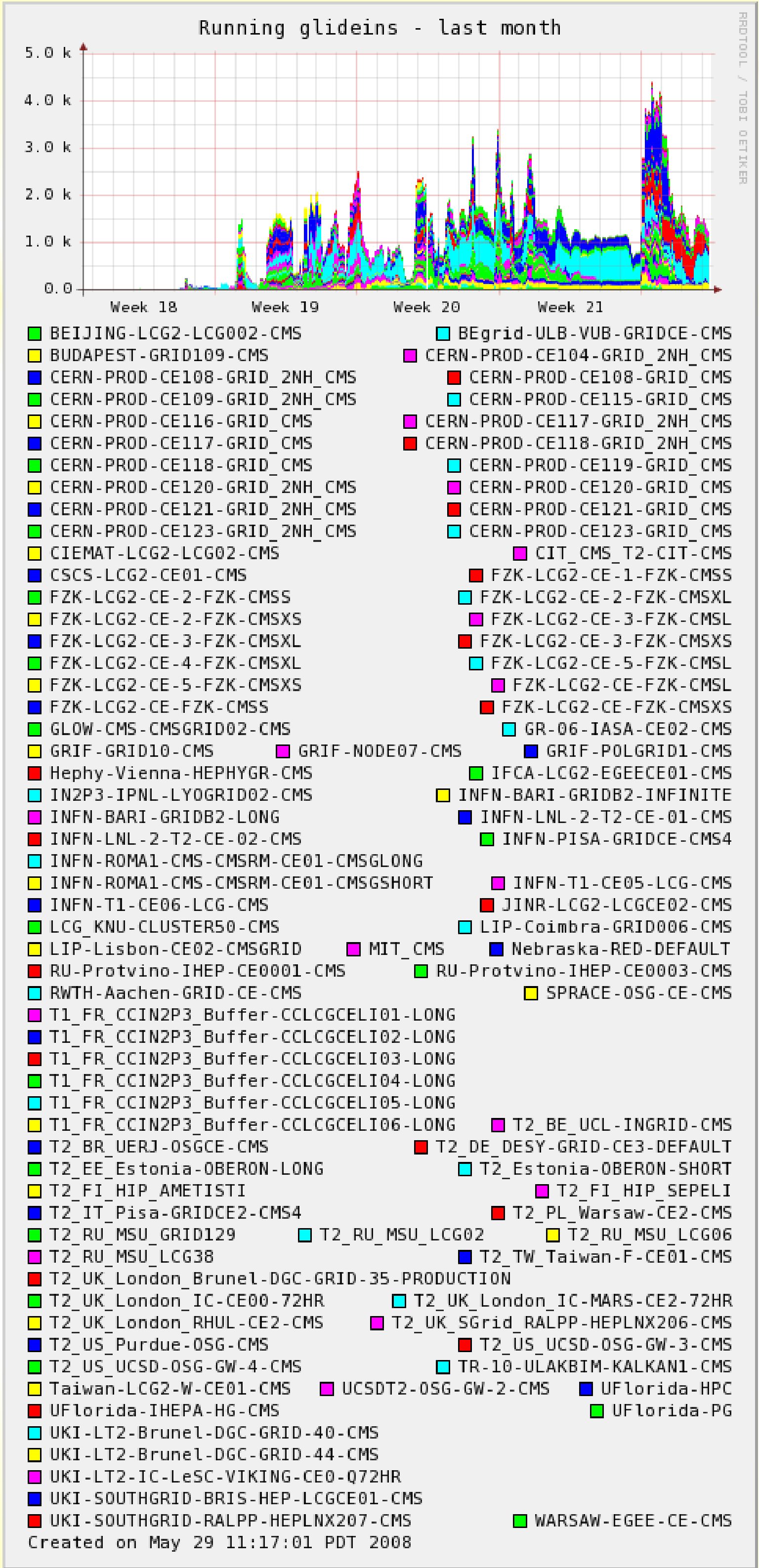
CMS @ FNAL setup

- 3 nodes used (+Grid worker nodes)
- No GCB (LAN)
- No gLExec (only production team)



CMS CCRC08 experience

- Using CRAB to submit to the local schedd(s)
- Submitting to 40 T2s
  - All over the world
  - OSG, EGEE and Nordugrid (a first for CMS)
- Ran 300k jobs over 4 weeks
  - Mix of CPU intensive and IO intensive jobs

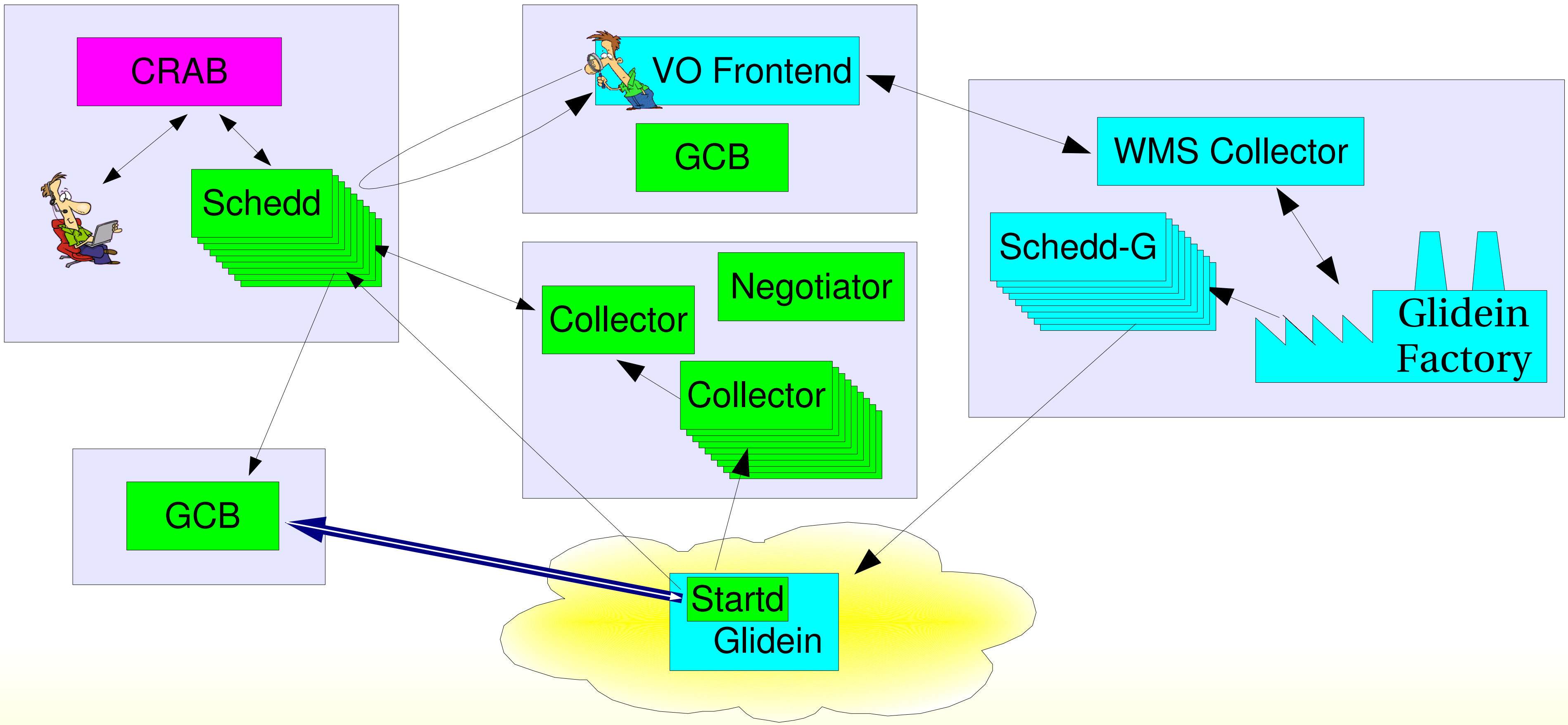


CMS CCRC08 experience (2)

- Latencies have bitten us
  - Condor uses blocking connections for security handshake
    - Condor working on fix
  - For CCRC solved by using multiple condor daemons
    - Hierarchy of collectors
    - Multiple schedds
- Still very successful
  - CMS pleased with the experience

CMS CCRC08 setup

- 5 nodes used (+Grid worker nodes)
- No gLExec (Only one test user)



CMS glidein plans

- Production over all T1s using glideinWMS should start soon (from FNAL)
  - Prototype in place
  - Need to sort out operational issues
- UCSD offered to host an analysis service
  - Serving physicists
  - Using the CRABServer
  - Using gLExec
  - Expected to be setup over the summer

Conclusions

- Bare-bones Grid difficult to use
  - Glideins can hide the Grid complexity and make it look as a uniform computing pool
- CMS has used glideinWMS for the past 6 months
  - Great success at FNAL
  - Good results in tests over T1s and T2s

Backup Slides

glideinWMS contact info

GlideinWMS home page:

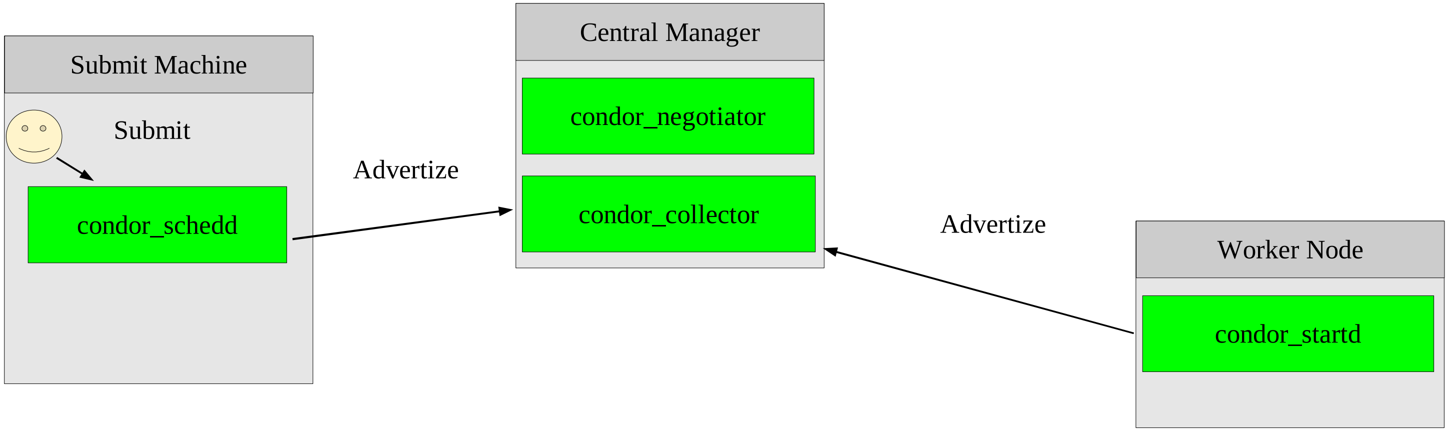
http://www.uscms.org/SoftwareComputing/Grid/WMS/glideinWMS/

Condor home page:

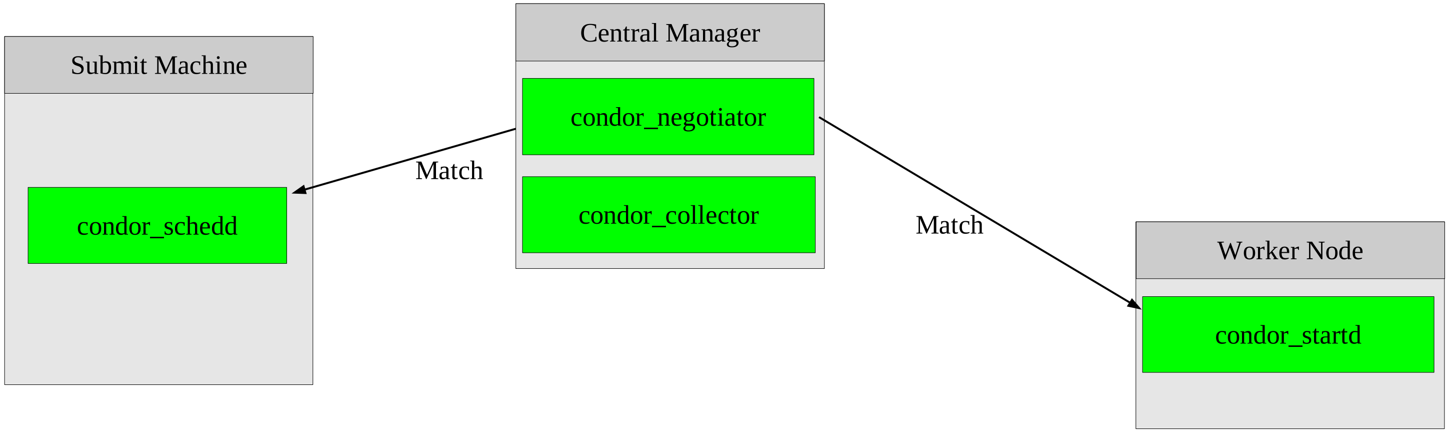
http://www.cs.wisc.edu/condor/

email: sfiligoi@fnal.gov

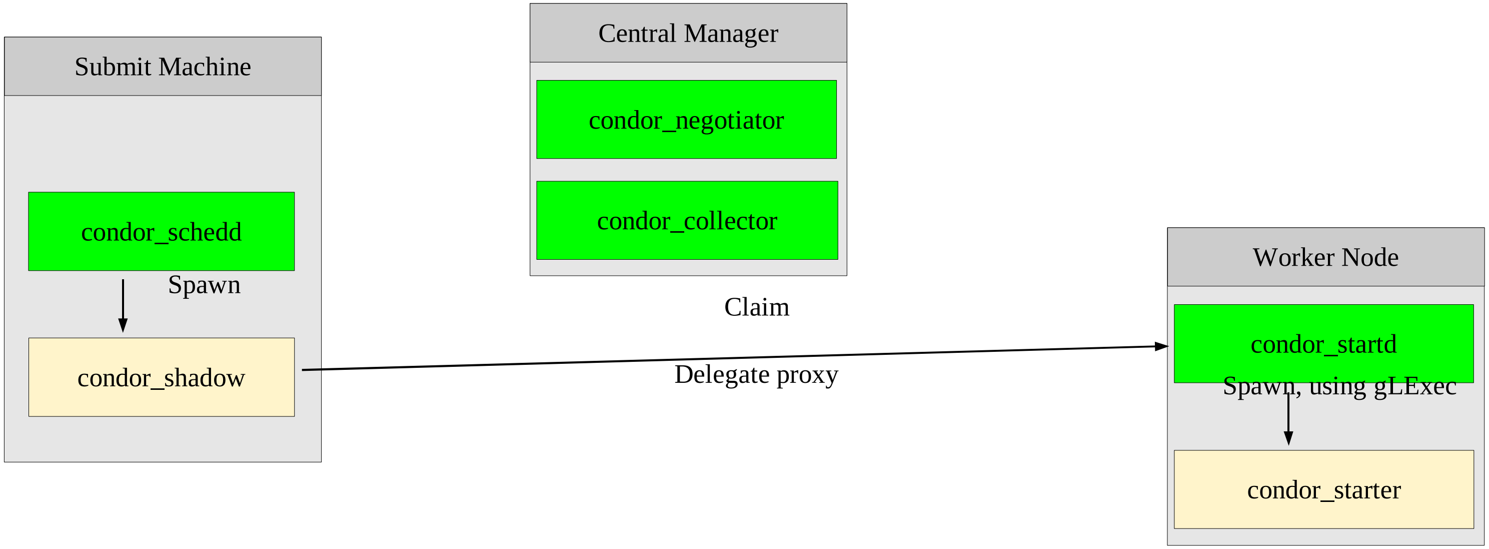
Condor Internals



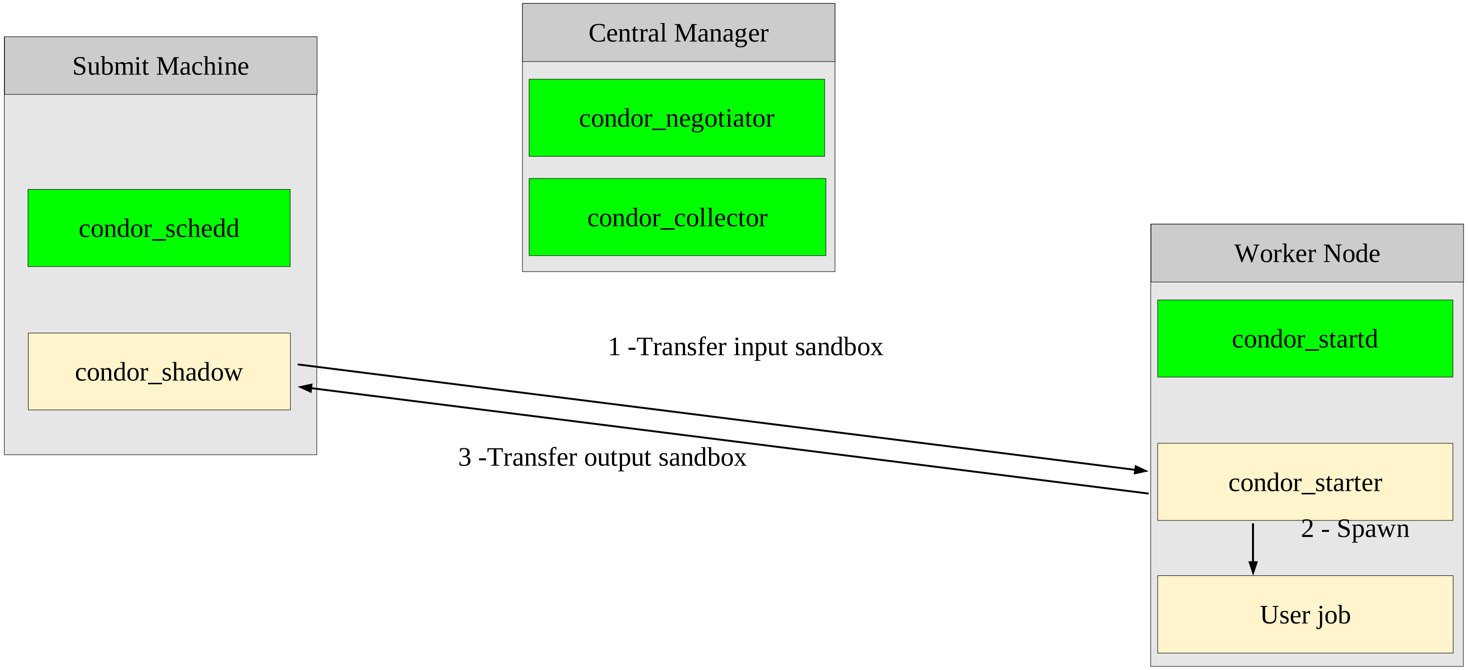
Condor internals



Condor internals



Condor Internals


Submit Machine
condor_schedd
condor_shadow
Central Manager
condor_negotiator
condor_collector
Worker Node
condor_startd
condor_starter
2 - Spawn
User job
1 -Transfer input sandbox
3 -Transfer output sandbox